图1显示了我们提出的方法的总体框架，该框架包括四个部分：1）从音频中提取对数幅度mel-频谱图； 2）用一维ConvNet转换频谱特征； 3）将时间相关性与RNN-LSTM结合在一起；以及4 ）通过后处理来确定音频事件的存在和开始时间。

将一个窗口应用于大小为46 ms的音频信号，该窗口与窗口的一半大小重叠。 我们还在每个帧的频谱上应用了128个滤波器组，并对幅度取对数。 mel频谱图被分为一个具有时间步长（τ）的块，并馈入我们的网络。

许多使用2D输入功能的音频内容分析研究，例如频谱图，梅尔频谱图和梅尔频率倒谱系数（MFCC），都应用了2D ConvNet [13]，[14]，这经常用于图像内容分析。它着重于音频特征的频谱和时间局部性，以提取有意义的信息。但是，基于2D ConvNet的方法以块级而非帧级分析音频。由于精确的估计开始时间对于该任务是必要的，因此我们应用频谱侧的一维ConvNet来进行帧级研究。 1D ConvNet步骤包括1D卷积层，批处理规范化（BN）[15]过程和池化层。图2显示了1D卷积层和最大合并层的概念。卷积层的大小设置为32，总共使用128个过滤器。因此，当梅尔频谱图（128频点）的单个帧适合于1D卷积层时，将产生128个输出，每个包含97（128-32 + 1）个元素。在下一步中，将BN应用于要素图输出，以便它们将主题保持在0附近，将标准偏差保持在1附近。之后，整流线性单元（ReLU）[16]被应用到sansanda激活函数。最后，将97大小的最大池应用于每个输出以提取代表值。在ConvNet的末尾也将Dropout的值设置为0.3，以防止过度拟合。

RNN已被证明是用于识别诸如语音识别[17]和手写识别[18]之类的顺序信息的强大模型。特别地，RNN-LSTM是一种众所周知的深度学习模型，可防止梯度消失，而梯度消失会干扰长期序列学习[19]。因此，我们在这里使用RNN-LSTM合并提取特征的时间依赖性。在这里，我们使用两个RNN层，每个层包含128个LSTM单元。与使用前向或双向RNN-LSTM的常规研究不同，我们使用单向后向RNN-LSTM。这是因为与事件之前的信息相比，事件发生后的信息对于准确的事件检测似乎更为重要。根据我们的实验，这种单向向后分析显示出比其他方法更好的性能。图3显示了RNNLSTM步骤内部的处理结构。从ConvNet中提取的特征（xt，xt + 1，...，xt +τ-1）被馈送到将其通过各层的网络中。注意，针对每个帧获得128维输出向量（zt，zt + 1，...，zt +τ-1）。我们使用双曲正切（tanh）作为激活函数，并对所有RNN-LSTM层应用0.3的下降率。

来自RNN-LSTM层的功能转变为包含128个隐藏单元的非完全连接层（FC）。与之前的1D ConvNet步骤相似，BN和ReLU分别用作规范化函数和激活函数。然后将更新的特征转发到具有一个S形单位的时间分布的输出层，其输出表示存在目标声音事件的概率。结果，在时间步长期间为梅尔谱图的每个帧计算了概率值。为了在测试阶段获得整个音频片段的概率序列，采用了滑动集成法。由于使用我们训练有素的模型在每个数据块中计算了概率值，因此该方法通过将预测数据块以一帧的跳数（23毫秒）滑动并平均存在该值的索引的概率，从而将整个概率组合在一起。图4显示了此方法。图5显示了确定事件存在并根据概率序列预测事件的开始的示例。为了确定声音事件的存在，使用具有经验假设的硬阈值方案。如果概率序列中的最大值大于0.8（“枪声”为0.5），则音频片段被视为包含目标事件。为了找到声音事件的开始时间，我们从最大值中的前50个帧（“婴儿哭声”为200）中选择大于0.5的值作为第一个索引。

对于DCASE 2017的任务2，提供了“ TUT Rare Sound Events 2017”数据集，该数据集包含每个目标类别的隔离声音事件和日常声学场景的记录作为背景。在数据集中，考虑了三个目标声音事件：“婴儿哭泣”，“玻璃破碎”和“枪声”。背景音频集包含来自15种不同音频场景的录音，这是“ TUT Acoustic Sc​​enes 2016”数据集的一部分。还提供了用于创建不同事件与背景的组合的源代码以及录音。使用该代码，我们可以生成具有不同参数的训练数据，例如混合数，事件对背景比率（EBR）和事件发生概率。混合注释也将自动生成，包括目标事件的名称及其时间位置。我们创建了4组混合物（S1，S2，S3，S4）。每组包含15,000个音频片段（每个事件类别5,000个），它们的EBR为6,0,6dB，事件发生概率为0.5。所有混音都被创建为具有44,100 Hz和24位的30秒单声道音频。为了进行训练，这些混合物随机分为8组和2组。在测试过程中使用包含1,500个音频剪辑（每个事件类别500个）的预组合测试集。

在训练阶段，在将输入块输入模型并转换为概率值之后，使用二进制交叉熵作为损失函数来计算预测值和正确值（0或1）之间的误差。 为了优化损耗，我们将自适应动量（Adam）用作优化器，并将小批量的大小设置为256。学习率最初设置为0.001，并在每个时期衰减，学习率的衰减系数为0.01。 当验证损失停止改善10个纪元时，将尽早停止学习。

<https://github.com/ksanjeevan/crnn-audio-classification>